2019

A theoretical framework for research on readmission risk prediction

Isabella Eigner

Freimut Bodendorf

Nilmini Wickramasinghe

32ND BLED ECONFERENCE
HUMANIZING TECHNOLOGY FOR A SUSTAINABLE SOCIETY
JUNE 16 – 19, 2019, BLED, SLOVENIA, CONFERENCE PROCEEDINGS
A. Pucihar (ed. et al.)

University of Maribor Press

A theoretical framework for research on readmission risk prediction

ISABELLA EIGNER, FREIMUT BODENDORF &
NILMINI WICKRAMASINGHE

Abstract On the one hand, predictive analytics is an important field of research in Information Systems (IS); however, research on predictive analytics in healthcare is still scarce in IS literature. One area where predictive analytics can be of great benefit is with regard to unplanned readmissions. While a number of studies on readmission prediction already exists in related research areas, there are few guidelines to date on how to conduct such analytics projects. To address this gap the paper presents the general process to develop empirical models by Shmueli and Koppius (2011) and extends this to the specific requirements of readmission risk prediction. Based on a systematic literature review, the resulting process defines important aspects of readmission prediction. It also structures relevant questions and tasks that need to be taken care of in this context. This extension of the guidelines by Shmueli and Koppius (2011) provides a best practice as well as a template that can be used in future studies on readmission risk prediction, thus allowing for more comparable results across various research fields.

Keywords: • Readmissions • Predictive analytics • Prediction process • Healthcare • Framework •

CORRESPONDENCE ADDRESS: Isabella Eigner, FAU Erlangen-Nuremberg, Germany e-mail: isabella.eigner@fau.de. Freimut Bodendorf, FAU Erlangen-Nuremberg, Germany e-mail: freimut.bodendorf@fau.de. Nilmini Wickramasinghe, Swinburne University of Technology, Hawthorn, Australia, nilmini.work@gmail.com.

DOI https://doi.org/10.18690/978-961-286-280-0.21 ISBN 978-961-286-280-0
Dostopno na: http://press.um.si

1 Introduction

Hospital readmissions, especially unplanned readmissions are an important quality measure in healthcare, as they can indicate issues around treatments, rehabilitation and/or discharge management. Moreover, readmissions are often associated with increased costs resulting from penalties and regulations enforced by policy makers and insurers. At the same time, the increasing availability of healthcare data leads to an uptake in predictive analytics research conducted in the healthcare sector. The identification of patients at high risk of readmission is a significant issue in this context. The main motivation behind this research area is to identify patterns that can help to unravel high-risk patients to allow for timely interventions. The starting point of these interventions lies in the screening of individuals at high risk of discharge failure (Scott, 2010). By identifying high-risk patients, hospital resources can be allocated accordingly and interventions and discharge planning can be adapted. Multiple factors associated with a higher risk of readmission have been identified in research, including health factors (e.g., co-morbidities (Kumar et al., 2017; van Walraven, Bennett, Jennings, Austin, & Forster, 2011), social factors (e.g., marital status (Hasan et al., 2010)), clinical factors (e.g., hospital utilization (Shadmi et al., 2015)), length of stay (Heggestad, 2002)) or effective discharge management (Ohta, Mola, Rosenfeld, & Ford, 2016).

Determining the risk of readmission is an imperative and highly complicated task, relying on different risk factors for various health conditions. While some studies propose general risk scores (Donzé, Aujesky, Williams, & Schnipper, 2013; van Walraven et al., 2010) applicable for all kinds of diseases, research shows significant variation in risk factors for different health conditions. Thus, to be able to accurately predict patients at high risk of readmission, individual prediction models for different health conditions should be preferred. Even though there are a number of studies dealing with this phenomenon, currently no theoretical framework exists to guide these kinds of research projects. This leads to the issue that studies on readmission risk prediction often disregard key characteristics for this prediction task. Also, results from different studies are often difficult to compare and thus unsuitable to generalize best practices. This study proposes a theoretical framework to guide studies on readmission risk prediction by providing a structured overview of relevant definitions, tasks and questions that need to be taken care of in this context. To identify these steps

previous studies are analysed to identify project characteristics specifically for hospital readmission prediction.

2 Theoretical and conceptual background

2.1 Readmissions in hospitals

While there is no standard definition for readmissions available, they can be broadly described as "a second admission to a hospital within a specified period after a primary or index admission" (Kristensen, Bech, & Quentin, 2015, p. 265). For each healthcare system, criteria concerning the index admission and the second admission to account as a readmission as well as the considered time frame, have to be defined. These criteria can include clinical characteristics (e.g., diagnosis), demographics (e.g., patient age), type of the admission (e.g., elective or emergency) or the treatment facility (Kristensen et al., 2015). To determine the applicable time frame, readmission days are counted from the discharge date of the index admission until the admission date of the second admission. Consequently, a readmission is defined by the relation between two admissions and the time frame in between. There is no international consensus considering the specified period between admissions. The time frame varies among studies from 14-day to 4-year with the most common being 30-day readmissions (Kansagara et al., 2011).

2.2 Predictive analytics

Predictive analytics methods are used in a variety of application fields to extract patterns from historical data to create empirical predictions as well as methods for assessing the quality of those predictions in practice (Shmueli & Koppius, 2011). Predictive analytics are part of data mining, which aims at deriving models that can e.g., use patient-specific information to predict a specific outcome. As opposed to descriptive models that aim to identify human-interpretable patterns and associations in existing data based on pre-defined attributes, predictive analytics tries to foresee outcomes or classifications for new input data using a special response variable, thus the classification (Bellazzi & Zupan, 2008).

Shmueli and Koppius (2011) present a general approach for conducting predictive analyses. They postulate that in general, predictive analyses consist of two components: First, the empirical predictive model, such as statistical methods or data mining algorithms and second, methods that evaluate the predictive power of a model. The latter refers to the ability of a predictive model to accurately represent new observations. The explanatory power, in turn, is related to the strength of the association induced by the statistical model (Shmueli & Koppius, 2011). Figure 1 illustrates the general process steps, which are carried out for the creation of all empirical models. The individual tasks in this process, however, differ extensively when developing an explanatory or predictive model. For example, while explanatory models investigate the explanatory power of their identified relationships (e.g., theoretical coherence, strength-of-fit, statistical significance), predictive models assess the predictive accuracy of a model, e.g., using cross-validation or split-validation measures. The individual modelling steps as proposed by Shmueli and Koppius (2011) guide the development of the readmission prediction framework presented in this paper.

Figure 1: Process to build an empirical model (Shmueli & Koppius, 2011)

2.3 Imbalanced data

A major concern in predicting readmissions is the occurrence of imbalanced data. Imbalanced data, also known as skewed data, has a strong unequal distribution of the minority and majority classes (Sun, Wong, & Kamel, 2009). In the case of hospital readmissions, this is especially true for unplanned readmissions, as rates usually vary between 1.1 to 6.7 % (Kreuninger et al., 2018). The main issue with handling imbalanced data is that traditional classifiers tend to perform best with an equal class distribution while the relevant information from the minority class might be overlooked with regards to the majority class (Sun et al., 2009, 2009). There are a number of different approaches to handle imbalanced data (Nitesh Chawla, 2005; Galar, Fernandez, Barrenechea, Bustince, & Herrera, 2012; He & Garcia, 2009; Kotsiantis, Kanellopoulos, & Pintelas, 2006; Longadge & Dongre,

2013; Sun et al., 2009), the most popular being sampling or ensemble techniques (Haixiang et al., 2017).

Sampling

Two main sampling approaches can be differentiated, namely oversampling and undersampling. Undersampling reduces the entities from the majority class, while oversampling creates additional entities of the minority class (Galar et al., 2012; Kotsiantis et al., 2006). A variety of sampling approaches are available to reach this goal, the most prominent being random over- and undersampling, informed undersampling, synthetic minority oversampling (SMOTE), adaptive synthetic sampling, sampling with data cleaning, and cluster-based sampling methods (He & Garcia, 2009). From the variety of over- and undersampling methods presented in literature (Galar et al., 2012; Haixiang et al., 2017), random undersampling (RUS) is still one of the most commonly applied undersampling techniques (Haixiang et al., 2017). In RUS, entities of the majority class are randomly removed to reduce the data imbalance (Galar et al., 2012). The most commonly used oversampling technique is SMOTE and its derivations (Haixiang et al., 2017). The SMOTE process is introduced by Chawla et al. (2011; 2003). For each entity of the minority class, the k-nearest neighbours are identified; after this, a distance vector from the minority entity to its neighbours is calculated. By randomly multiplying the vector with a number between 0 and 1, SMOTE creates a new data entity, which is added to the training data.

Ensemble learning

Hybrid methods of predicting imbalanced data include cost-sensitive learning and ensemble learning. Cost-sensitive learning follows the approach of manipulating the algorithm to weight the minority class higher and improve classifier performance. Cost-sensitive approaches have the downside that the actual costs of misclassification must be known (Sun et al., 2009). Another issue in readmission prediction as pointed out by Kansagara et al. (2011) is the poor performance of individual classifiers. Ensemble methods counter this issue by combining multiple classifiers into one classification system to produce a higher accuracy than achieved by its individual components (Galar et al., 2012). Ensemble learning can either be performed by combining different classifiers or by applying variations of the same classifier (Haixiang et al., 2017). Two main

approaches that can be differentiated are bagging and boosting. Bagging, which is short for bootstrapped aggregating, is introduced by Breiman (1996) and combines several base classifiers into one classifier. In the first step, data subsets are sampled from the training data. The bagging approach bootstraps the data to create several different bags. Bootstrapping means that random samples are added to the subset until the subsets have the same number of entities as the training data. This leads to intended duplicates in the subsets. Next, for each of the bags, the base classifier is trained and applied to the application set. Subsequently, the differently trained classifiers vote as to which class a new entity belongs, and a majority vote of the classifiers determines in which class the observation fits best. A prominent bagging method are RandomForests, which combine individual decision trees into a single classifier. In boosting, the training set is again split into k subsets. The model building, however, is done sequentially as opposed to the independent training for bagging models. Here, a weight is set for each data element, where misclassified examples increase their weight for the subsequent training round. In addition, a weight is set for each classifier dependent on its individual error rate. Thus, a weighted vote from all classifiers is used for the prediction of a new example (Quinlan, 1996). The most prominent boosting method, AdaBoost (adaptive boosting) (Freund & Schapire, 1997) is based on the principle of boosting introduced by Schapire (1990) and uses the base principle of improving the algorithm in every iteration to achieve a higher performance. Here, the base classifier is applied to the entire training data set. Next, AdaBoost calculates the error rate for each individual sample and adds it to the data. In the next iteration, the algorithm selects the training data by considering the assigned weight to give misclassified samples higher attention. After each iteration, AdaBoost weights the models according to accuracy.

3 Framework development

3.1 Goal definition and study design

As a first step in any prediction project, the analysis goal has to be defined. While the main objective is to predict patients at risk of readmission to the hospital, the specific terms and criteria to successfully reach this goal need to be defined, namely the *type of prediction*, the interpretation of a *high-risk patient* as well as the parameters for an episode to count as a *readmission*.

Prediction: In supervised learning, two types of prediction tasks can be differentiated, namely classification and regression. Classification aims at predicting discrete values, i.e. predefined categories or classes, whereas regression provides continuous values. In the task of readmission prediction, a categorical value, hence a classification approach, is required. At the highest level, a dichotomous differentiation between readmitted and non-readmitted patients is chosen. If necessary, the classes can be extended to further distinguish the time of readmission (e.g., early versus late readmissions) in a specified time frame, they can be separated by the reason for readmission (e.g., complications or corrections) or the level of risk (e.g., low risk, medium risk, high risk). The main issue for each of these cases is the prior classification of examples in the historical dataset that has to be aligned with the goal of the analysis task. In the case at hand, the main goal is to find out, whether a patient will be readmitted or not. Thus, a binary variable reflecting either 1 (readmission) or 0 (no readmission) is chosen as the classification target.

High-risk: The binary distinction can further be extended by considering the probability of class memberships. This way, prediction models cannot only specify, whether a patient belongs to the predicted readmission group or not, but also the probability of belonging to a group can be determined. The lower the threshold for a required class membership is set, the more risk patients can be identified. On the other hand, this also increases the likelihood of false positives. If the costs for a false positive prediction or a false negative prediction are known, weights can be specified accordingly. The concrete value of wrong predictions, however, is difficult to determine and poses a major challenge in readmission prediction. Costs for a prolonged length of stay or intervention programs can be used as approximations (Jamei, Nisnevich, Wetchler, Sudat, & Liu, 2017).

Readmission: Another issue in readmission prediction lies in the basic definition of the readmission episode itself. Readmissions are commonly differentiated between planned or unplanned readmissions and related or unrelated to the index admission (AHA, 2011). While the identification and prediction of readmissions should primarily focus on unplanned, related readmissions, it is often difficult to assess the relationship between admissions. Also, planned readmissions are often not documented within hospitals and therefore exacerbate the distinction of unplanned readmissions. Besides the admission intent, some studies also differentiate between avoidable and

unavoidable readmissions (van Walraven et al., 2011; van Walraven, Wong, & Forster, 2012). The proportion of avoidable readmissions in that context and the underlying criteria to determine whether they are indeed avoidable varies strongly between studies. For example, van Walraven et al. (2011) suggest a median proportion of around 27 % of readmissions to be avoidable, or similarly van Galen et al. (2017) propose 27-28 % be at least predictable.

To specify which episodes qualify for this definition, a variety of factors, including the timespan between admissions and the reasons for readmission have to be clarified. The timeframe can be selected based on regulations at a country or hospital-level or adhere to protocols by insurers. The reasons for readmission to be related to the index admission are highly dependent on the episodes under study. If certain diagnoses or procedures are investigated, the most common diagnoses for readmissions can be identified apriori and categorised into the presented scheme for readmissions (AHA, 2011). This task requires sufficient domain knowledge to undertake the classification for a specific procedure or diagnosis group. Alternatively, existing guidelines or regulations by insurers or governments can also be used.

3.2 Results

The data preparation process covers various steps of cleaning, visualising and reducing the available dataset in order to be suitable for the subsequent analysis. This includes dealing with missing and inconsistent data as well as creating and selecting appropriate features. To get a better understanding of the underlying data and identify noise, exploratory analysis and simple visualizations of the dataset are conducted. Figure 2 gives an overview of the individual steps that are taken to develop the appropriate feature sets in the following sections.

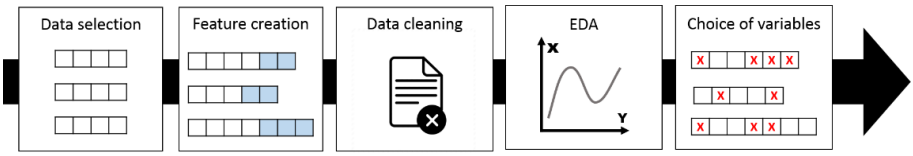

Figure 2: Data preparation steps

Data selection

As a first step, data is filtered to only include relevant admissions for the prediction task. It is imperative that the prediction model is trained on the data of the admission episodes that might have led to a readmission, not on the readmission episodes. The following criteria are important for each episode to remove irrelevant data points accordingly:

- The patient is admitted to acute care.
- The patient did not die during or after the hospital stay.
- The patient did not leave the hospital at his/her own risk.

Feature creation

To complement the data set with further relevant attributes, the availability of the identified risk factors from previous studies is assessed for each procedure group. Based on the insights from systematic reviews by Kansagara et al. (2011) and Zhou et al. (2016) relevant attributes for readmission risk prediction from previous studies can be analysed and, if applicable, integrated into the dataset. Furthermore, if no studies on predictive models are available for the diagnosis or procedure under study, explanatory models can also provide an indication of relevant risk factors.

Data cleaning

The term data cleaning describes the process of detecting and removing data errors and inconsistencies. Unclean data can either occur on attribute, record, record type, or source level. According to Rahm (2000) errors can appear on a schematic or at an instance level.

Schematic errors can consist of the following:

- illegal values in attributes (e.g., a BMI of 0),
- inconsistencies on record level (e.g., between age and date of birth),
- record type errors, such as uniqueness violations (e.g., multiple uses of patient or episode IDs), or
- referential integrity violations (e.g., missing descriptions of diagnosis codes).

On an instance level,

- missing values,
- misspellings,
- abbreviations or non-defined codes,
- embedded values (i.e., multiple attributes in one column), or
- misfielded values (e.g., age in the date of birth column) can occur.

Duplicate records or varying value representations (e.g., data types) also affect the integrity of the data set (Rahm & Do, 2000). According to Chen et al. (2014), completeness, accuracy, and timeliness of data are the major factors for data quality specifically in health information systems. To identify errors, data profiling can be performed, which provides metadata to discover errors in the data.

Missing values can be handled in different ways, where entities can either be deleted, missing values can be imputed or the missing values can present knowledge themselves (Grzyb et al., 2017). If missing values don't indicate additional insights, attributes with too many missing values are not taken into further consideration. Also, attributes contributing low or now information are identified by calculating the variance of each variable. Attributes with a variance lower than a predefined threshold can be excluded from the dataset.

Exploratory data analysis

The goal of the exploratory data analysis (EDA) is to analyse the dataset visually and numerically to ensure that the data is suitable for the prediction model. In addition, dimensions are systematically reduced in this step as too many predictors can introduce noise and thus decrease the performance of a prediction model. Depending on the type of the attribute under study, different graphical representations can be used to gain insights into the analysed records. For univariate and bivariate data (e.g., gender), simple plots, such as histograms or scatterplots can be used. The numerical distribution gives an insight into how the two cohorts differ.

Choice of variables

After reducing dimensions, the next step is to select which variables to use for the prediction models. To this end, the variables must have a measurement quality, which means variables that do not assist in predicting unplanned readmissions are not relevant for the model. A feature is seen as beneficial if it is correlated with the prediction flag but is not redundant to any other relevant feature (Yu & Liu, 2003). This means that the variables must have the ability to predict readmissions while not being highly correlated with each other. Since variables with correlations above 0.70 are seen as highly correlated (Asuero, Sayago, & González, 2007), features with a correlation above 0.70 can be removed.

An additional aspect that distinguishes prediction models from explanatory models is the time of data availability. While explanatory models can utilize all data that is available to identify relationships a posteriori, prediction models need to be based on data that is available at the time of prediction (Shmueli & Koppius, 2011). As the prediction models are usually utilized before patient discharge, only attributes that are available before a patient leaves the hospital can be considered.

3.3 Model development

According to a systematic review by Artetxe et al. (2018) on predictive models for hospital readmission risk, machine learning methods can improve the prediction ability over traditional statistical approaches. Such contributions to this academic field are aimed at first aligning complex and sensitive information across multiple sources, using, among others, administrative, insurance, clinical, and government registry data. This information is thereafter used to identify patients in need of additional healthcare resources by means of various intervention methods (Billings, Georghiou, Blunt, & Bardsley, 2013). The model development is split into several steps (cf. Figure 3) and is tightly connected to the internal evaluation and optimization of a prediction model.

Process step	Input	Output
Split dataset	Dataset	*Training set* *Validation set*
Sampling	Training set	*Sampled training set*
Feature selection	Sampled training set	*Sampled training set with relevant attributes*
Hyperparameter tuning	Sampled training set with relevant attributes	*Optimal hyperparameters*
Model building	Sampled training set with relevant attributes + Optimal hyperparameters	*Prediction model*

Figure 3: Model development process

Split the dataset

As a first step, the prepared dataset is split into a training and a validation set. The training set is further used to train, test and optimize the models, while the validation set is used in the very last step to evaluate and compare the predictive performance of the final models. The data is split in a stratified fashion, thus the distribution of readmitted and non-readmitted patients is equal in both datasets. A major issue in predictive analytics is overfitting, which refers to a model that fits the training data perfectly, but fails to generalize in order to correctly predict new examples. Different strategies can be applied during model training to avoid and test if a model overfits, namely hold-out validation and cross-validation. To perform these validations, the data is split into three subsets, a training set and a validation set for cross-validation or hold-out validation and a test set for final evaluation. Depending on the evaluation strategy, these sets are created and used in different manners.

- **Training set:** This subset is used to fit the model, i.e., derive the relationship between the input variables and the target class.

- **Validation set:** Next, the developed model is tested on unseen data, where the predicted values are compared with the real class membership to determine the error rate of the predictions.
- **Test set:** The test set is used in the last step to evaluate the final model that is built on the full dataset (training + validation) given the optimal hyperparameters previously determined by the training and validation data.

For both approaches, a test set is omitted for final testing of the developed model. The training and validation of the model, however, differs. In hold-out validation, for each parameter setting, the model is only trained once on the training set and then applied to the validation set. When the best parameter setting is found through this approach, the final model is again trained on the entire dataset (training + validation data) and then evaluated using the test set. In cross-validation, on the other hand, the data is split into k subsets, where k equals a positive integer. Next, the model is trained on k-1 subsets and validated on the remaining subset. This is repeated until every subset has been used as a training and validation set (cf. Figure 4). A special form of cross-validation, termed leave-one-out cross-validation (LOOCV), splits the data into k subset, where k equals the number of examples in the dataset. Thus, each data point is used on its own to evaluate the model that is built on the remaining dataset. This approach, however, gets extremely cost-intensive with regards to computing time the bigger the data set. While hold-out validation requires less computing time as the model only has to be trained once, sampling of the training and test set can lead to an unwanted bias. In cross-validation, on the other hand, each data point is used both as a training and a validation example, eliminating the sampling bias. Since computing time is not an imitating factor in this analysis and the size of the data sets is appropriate for cross-validation, this technique is used for evaluating the prediction models in the following sections.

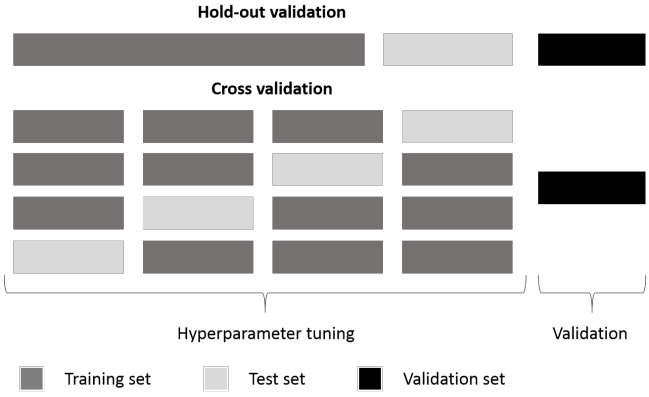

Figure 4: Hold-out validation versus cross-validation

Sampling

If the utilized algorithm doesn't support class weights, sampling can be performed on the training data set to handle an imbalanced class distribution. There is no clear suggestion, whether over- or undersampling performs better in a given prediction task, thus both approaches should be tested. In order to avoid shrinking the data set in the sampling process too extensively, the desired ratio between the minority and majority class can be specified.

Feature selection

Next, different feature selection approaches are performed for each classifier. In general, filter, wrapper and embedded methods can be distinguished (Guyon & Elisseeff, 2003). The main difference between these approaches lies in the point in time of feature selection with regards to the model development and evaluation (cf. Figure 5).

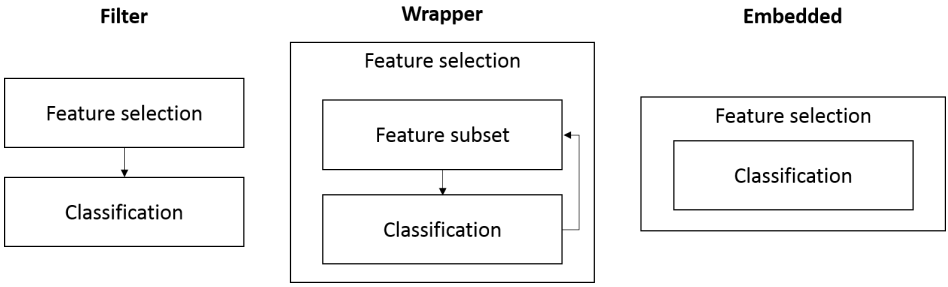

Figure 5: Feature selection approaches (cf. Suppers, van Gool, & Wessels, 2018, p. 7)

Filter methods clearly separate the feature selection and model building process. As a first step, attributes are chosen based on model-independent factors, such as variance or correlation thresholds. Wrapper methods, on the other hand, iteratively build and evaluate a model and adapt the feature set based on the results of the model evaluation until a certain threshold is reached. This adaption can be done by increasing or decreasing the number of features. In forwards selection, the initial feature set consists of one attribute that is consistently extended. The main issue with forwards selection is that features whose usefulness is dependent on other features ("feature synergy") might be lost (Kohavi & John, 1997). To overcome this issue, backwards elimination initially uses the entire feature set to build the classification model and attributes are iteratively removed. Recursive feature elimination (RFE) is a type of backwards selection, where the model is first trained on all features, which are then ranked based on their contribution to the prediction task. The lowest-ranking features are removed until the prediction accuracy of the model decreases. Lastly, embedded methods perform the feature selection task during model building. Decision trees are a prominent example of an embedded feature selection model, as the information gain of each attribute is used to choose the features for model building. Since KNN and NB can't consider varying importance of different features, the models are fitted on all attributes. L1 regularization (also termed "least absolute shrinkage and selection operator (LASSO)") also presents an embedded method for a linear regression that adds a penalty for overly complex models, i.e., the number of input factors. Since DT have an embedded method of feature selection based on the information value of attributes, RFE is performed with cross-validation (RFECV) for all other methods. In RFE,

attributes are continuously excluded from the data set based on their contribution to the prediction task.

Hyperparameter tuning

As a next step, hyperparameter tuning is performed where the classifier is fitted to the sampled training set with the remaining relevant attributes. Each model can be trained using a set of hyperparameters relevant for each algorithm. The hyperparameters determine various criteria on how a model is trained, the learning speed and the structure of the model. To identify the best combination of hyperparameters, different search strategies can be applied. With sufficient computing power, formerly popular manual "trial-and-error" settings can be neglected. Instead, parameter combinations can be tested within a given scope using search algorithms, such as random search or grid search. In random search, each parameter setting is sampled from a distribution over possible parameter values. On the other hand, grid search offers an exhaustive search in a specified scope parameter value. Research has shown that random search provides a more efficient way of identifying the optimal parameter setting with at least equally satisfying results (Bergstra & Bengio, 2012).

Model building

In the last step, the prediction model is built by training the classifier on the entire training and test data set using the identified hyperparameter combination. The resulting model can then be used for the final validation. Depending on the classifier, sample weights or embedded feature selection can be employed during model building. Otherwise, the over- or undersampled data is used to build the prediction model based on the previously identified relevant features.

3.4 Evaluation, validation, and model selection

In the last step, the prediction model is applied to the final test set. Thus, the model is tested on previously unseen data that hasn't been involved in the development process. A major issue in predictive analytics is overfitting, which refers to a model that fits the training data perfectly, but fails to generalize in order to correctly predict new examples. A popular strategy to test if a model overfits is to perform cross-validation. For this purpose, the data is split into

three subsets, a training set, test set and validation set. For the training and testing data sets, the data is split into k subsets, where k equals a positive integer. Next, the model is trained on k-1 subsets and tested on the remaining subset. This is repeated until every subset has been used as a training and testing set. While cross validation already aims to avoid overfitting of the model during training, it is argued that a final test on an unseen validation set should be performed in addition using data not present in the cross-validation (Ripley, 2009).

For evaluation, different metrics to investigate model performance are available. Since projects on readmission prediction usually concentrate on identifying as many risk patients as possible, the positive class should be focused on in the model evaluation. For this purpose, either the sensitivity or the F-2 score should be chosen as they put more emphasis on the positive class (cf. Table 1). Besides the resulting predictive performance stated by the evaluation metrics, model interpretability and computing time should also be considered for the final model selection.

Table 1: Evaluation metrics

Evaluation metric	Formula*
Accuracy	$\frac{TP + TN}{N}$
Sensitivity (Recall pos. class)	$\frac{TP}{TP + FN}$
Specificity (Recall neg. class)	$\frac{TN}{TN + FP}$
Precision	$\frac{TP}{TP + FP}$
F-Score	$(1+ \beta^2) * \frac{precision * recall}{(\beta^2 * precision) + recall}$

* TP = True Positives, TN = True Negatives, N = All examples, FN = False Negatives, FP = False Positives

Table 2 summarizes the results of this study by defining five main process steps that are further subcategorized in relevant tasks and questions that need to be answered in any readmission prediction project.

4 Discussion and Conclusions

This study set out to identify and unpack key issues around applying predictive analytics to healthcare especially in the area of hospital readmissions. In doing so the study has several contributions for theory and practice as follows: The proposed framework can be used to perform future studies on readmission risk prediction in a more systematic and guided way. Common mistakes in these kinds of projects can therefore be avoided and results are better comparable. Furthermore, this work extends the theoretical knowledge on predictive analytics based on Shmueli and Koppius (2011). In a next step, the proposed framework will be further tested and adapted by means of a systematic literature review on readmission risk prediction. Furthermore, an exemplary prediction project is conducted based on the presented guidelines to test its applicability in practice. For this purpose, episode data from an Australian hospital group is used to predict unplanned readmissions

Table 2: Framework for research on readmission risk prediction

Process Step	Main questions	Example
Goal definition	**Prediction:** What is the main purpose of the prediction?	*Define time of prediction, e.g., identify patients at risk for readmissions at admission, before or after discharge*
	High-risk: At what level should the readmission be predicted?	***Discrete:*** *Binary prediction (readmission / no readmission) or Multinominal prediction (e.g., high risk, medium risk, low risk)* ***Continous:*** *Risk probability (0 - 100 %)*
	Readmission: How is a readmission defined?	*Reason for readmission (procedure-specific or general)* *Timeframe of readmission (28-day, 30-day, 6 months, etc.)*
	Study design: When is the data collected?	*Retrospective study versus real-time*

Data collection and study design	**Data collection:** Which episodes should be excluded from the dataset?	*Patient is admitted to acute care* *Patient died before or after discharge* *Patient left the hospital against medical advice*
Data preparation	**Selection**: Which data points (episodes) are relevant for the context at hand?	*Focus on specific procedures, diagnoses, patient groups*
	Feature creation: Which additional attributes are potentially interesting?	*Create additional attributes from collected data that are not directly reported (e.g., from previous studies on predictive or explanatory models)*
	Cleaning: Which data points are usable for the prediction task?	*Missing values* *Outliers* *Low variance* *High correlation* *High cardinality*
	Exploratory data analysis: What does the population under study look like?	*E.g., use histograms or scatterplots to compare the distribution between two cohorts (readmission, no readmission)*
	Choice of variables: What data is available at the time of prediction?	*Depends on the prediction goal (at admission, before or after discharge)*
Model development	**Split dataset:** How does the data need to be split for evaluation?	*Training + test dataset (e.g., 80 %)* *(Final) validation dataset (e.g., 20 %)* *Cross-validation (during model training) versus holdout-validation*
	Sampling: Which sampling method should be applied to reduce the issue of imbalanced data?	*Methods that support class weights (e.g., SVM)* *Undersampling (e.g., Random Undersampling)* *Oversampling (e.g., SMOTE)* *Hybrid Sampling*

	Feature selection: Which attributes contribute to the predictive performance of a model?	*Filter methods (subsequent approach, e.g., variance threshold)* *Wrapper methods (iterative approach, e.g., backwards elimination)* *Embedded methods (integrated approach, e.g., decision trees)*
	Hyperparameter tuning: Which hyperparameter combination leads to the best predictive performance?	*Random search (specify the number of parameter combinations in a given range)* *Exhaustive search (test all parameter combinations in a given range, e.g., grid search)*
	Model building	*Build the model on the entire training + test dataset using the identified optimal hyperparameters* *Use sample weights and embedded feature selection (if applicable)*
Evaluation, validation, and model selection	**Evaluation:** Which evaluation measure should be chosen?	*Focus on readmission cohort: F2-score, Precision-recall curve, AUC*
	Validation: How well does the model perform on unseen data?	*Apply the model on the validation set*
	Interpretation: How can the final model be interpreted?	*Logistic regression: Odds ratio for each attribute* *Decision tree: Deduct rules from tree*
	Selection: What model should be selected for the final prediction task?	*Predictive performance, computing time, interpretability*

References

AHA. (2011). Examining the Drivers of Readmissions and Reducing Unnecessary Readmissions for Better Patient Care. Retrieved from https://www.aha.org/guidesreports/2018-02-09-examining-drivers-readmissions-and-reducing-unnecessary-readmissions

Artetxe, A., Beristain, A., & Graña, M. (2018). Predictive models for hospital readmission risk: A systematic review of methods. Computer Methods and Programs in Biomedicine, 164, 49–64. https://doi.org/10.1016/j.cmpb.2018.06.006

Asuero, A. G., Sayago, A., & González, A. G. (2007). The Correlation Coefficient: An Overview. Critical Reviews in Analytical Chemistry, 36(1), 41–59. https://doi.org/10.1080/10408340500526766

Bellazzi, R., & Zupan, B. (2008). Predictive data mining in clinical medicine: current issues and guidelines. International Journal of Medical Informatics, 77(2), 81–97. https://doi.org/10.1016/j.ijmedinf.2006.11.006

Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. Journal of Machine Learning Research, 13(Feb), 281–305. Retrieved from http://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf

Billings, J., Georghiou, T., Blunt, I., & Bardsley, M. (2013). Choosing a model to predict hospital admission: an observational study of new variants of predictive models for case finding. BMJ Open, 3(8), e003352. https://doi.org/10.1136/bmjopen-2013-003352

Breiman, L. (1996). Bagging Predictors. Machine Learning, 24(2), 123–140. https://doi.org/10.1023/A:1018054314350

Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. P. (2011). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. Advance online publication. https://doi.org/10.1613/jair.953

Chawla, Nitesh. (2005). Data Mining for Imbalanced Datasets: An Overview. In O. Z. Maimon & L. Rokach (Eds.), Data mining and knowledge discovery handbook (pp. 853–867). New York, NY: Springer. https://doi.org/10.1007/0-387-25465-X_40

Chawla, Nitesh, Lazarevic, A., Hall, Lawrence, & Bowyer, Kevin. (2003). Smoteboost: Improving Prediction of the Minority Class in Boosting. In N. Lavrač, D. Gamberger, L. Todorovski, & H. Blockeel (Eds.), Lecture Notes in Computer Science: Vol. 2838. Knowledge Discovery in Databases: PKDD 2003: 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, Cavtat-Dubrovnik, Croatia, September 22-26, 2003. Proceedings (Vol. 2838, pp. 107–119). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-39804-2_12

Chen, H., Hailey, D., Wang, N., & Yu, P. (2014). A Review of Data Quality Assessment Methods for Public Health Information Systems. International Journal of Environmental Research and Public Health, 11(5), 5170–5207. https://doi.org/10.3390/ijerph110505170

Donzé, J., Aujesky, D., Williams, D., & Schnipper, J. L. (2013). Potentially avoidable 30-day hospital readmissions in medical patients: derivation and validation of a prediction model. JAMA Internal Medicine, 173(8), 632–638. https://doi.org/10.1001/jamainternmed.2013.3023

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences, 55(1), 119–139. https://doi.org/10.1006/jcss.1997.1504

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42(4), 463–484. https://doi.org/10.1109/TSMCC.2011.2161285

Grzyb, M., Zhang, A., Good, C., Khalil, K., Guo, B., Tian, L., . . . Gu, Q. (2017). Multi-task cox proportional hazard model for predicting risk of unplanned hospital readmission. In 2017 Systems and Information Engineering Design Symposium (SIEDS) (pp. 265–270). IEEE. https://doi.org/10.1109/SIEDS.2017.7937729

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. The Journal of Machine Learning Research, 3, 1157–1182. Retrieved from http://dl.acm.org/ft_gateway.cfm?id=944968&type=pdf

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications, 73, 220–239. https://doi.org/10.1016/j.eswa.2016.12.035

Hasan, O., Meltzer, D. O., Shaykevich, S. A., Bell, Chaim M., Kaboli, P. J., Auerbach, A. D., . . . Schnipper, J. L. (2010). Hospital readmission in general medicine patients: a prediction model. Journal of General Internal Medicine, 25(3), 211–219. https://doi.org/10.1007/s11606-009-1196-1

He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Heggestad, T. (2002). Do Hospital Length of Stay and Staffing Ratio Affect Elderly Patients' Risk of Readmission? A Nation-wide Study of Norwegian Hospitals. Health Services Research, 37(3), 647–665. https://doi.org/10.1111/1475-6773.00042

Jamei, M., Nisnevich, A., Wetchler, E., Sudat, S., & Liu, E. (2017). Predicting all-cause risk of 30-day hospital readmission using artificial neural networks. PloS One, 12(7), e0181173. https://doi.org/10.1371/journal.pone.0181173

Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M., & Kripalani, S. (2011). Risk prediction models for hospital readmission: a systematic review. JAMA, 306(15), 1688–1698. https://doi.org/10.1001/jama.2011.1515

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. Artificial Intelligence, 97(1-2), 273–324. https://doi.org/10.1016/S0004-3702(97)00043-X

Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). Handling imbalanced datasets: A review. Retrieved from https://pdfs.semanticscholar.org/95df/dc02010b9c390878729f459893c2a5c0898f.pdf

Kreuninger, J. A., Cohen, S. L., Meurs, E. A. I. M., Cox, M., Vitonis, A., Jansen, F. W., & Einarsson, J. I. (2018). Trends in readmission rate by route of hysterectomy - a single-center experience. Acta Obstetricia Et Gynecologica Scandinavica, 97(3), 285–293. https://doi.org/10.1111/aogs.13270

Kristensen, S. R., Bech, M., & Quentin, W. (2015). A roadmap for comparing readmission policies with application to Denmark, England, Germany and the

United States. Health Policy (Amsterdam, Netherlands), 119(3), 264–273. https://doi.org/10.1016/j.healthpol.2014.12.009

Kumar, A., Karmarkar, A. M., Graham, J. E., Resnik, L., Tan, A., Deutsch, A., & Ottenbacher, K. J. (2017). Comorbidity Indices Versus Function as Potential Predictors of 30-Day Readmission in Older Patients Following Postacute Rehabilitation. The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences, 72(2), 223–228. https://doi.org/10.1093/gerona/glw148

Longadge, R., & Dongre, S. (2013). Class Imbalance Problem in Data Mining Review. Retrieved from http://arxiv.org/pdf/1305.1707

Ohta, B., Mola, A., Rosenfeld, P., & Ford, S. (2016). Early Discharge Planning and Improved Care Transitions: Pre-Admission Assessment for Readmission Risk in an Elective Orthopedic and Cardiovascular Surgical Population. International Journal of Integrated Care, 16(2), 10. https://doi.org/10.5334/ijic.2260

Quinlan, J. R. (1996). Bagging, boosting, and C4.S: AAAI Press.

Rahm, E., & Do, H. H. (2000). Data Cleaning: Problems and Current Approaches. In Zinc Industry (Introduction/page i-Introduction/page ii). S.l.: Woodhead Pub. https://doi.org/10.1016/B978-1-85573-345-9.50005-4

Ripley, B. D. (2009). Pattern recognition and neural networks (Reprinted.). Cambridge [etc.]: Cambridge University Press.

Schapire, R. E. (1990). The strength of weak learnability. Machine Learning, 5(2), 197–227. https://doi.org/10.1007/BF00116037

Scott, I. A. (2010). Preventing the rebound: improving care transition in hospital discharge processes. Australian Health Review : a Publication of the Australian Hospital Association, 34(4), 445–451. https://doi.org/10.1071/AH09777

Shadmi, E., Flaks-Manov, N., Hoshen, M., Goldman, O., Bitterman, H., & Balicer, R. D. (2015). Predicting 30-day readmissions with preadmission electronic health record data. Medical Care, 53(3), 283–289. https://doi.org/10.1097/MLR.0000000000000315

Shmueli, G., & Koppius, O. (2011). Predictive Analytics in Information Systems Research. MIS Q, 35(3), 553–572. https://doi.org/10.2139/ssrn.1606674

Sun, Y., Wong, A., & Kamel, M. (2009). Classification of imbalanced data: A review. International Journal of Pattern Recognition and Artificial Intelligence, 23(04), 687–719. https://doi.org/10.1142/S0218001409007326

Van Galen, L. S., Brabrand, M., Cooksley, T., van de Ven, P. M., Merten, H., So, R. K., . . . Nanayakkara, P. W. (2017). Patients' and providers' perceptions of the preventability of hospital readmission: a prospective, observational study in four European countries. BMJ Quality & Safety. Advance online publication. https://doi.org/10.1136/bmjqs-2017-006645

Van Walraven, C., Bennett, C., Jennings, A., Austin, P. C., & Forster, A. J. (2011). Proportion of hospital readmissions deemed avoidable: a systematic review. CMAJ : Canadian Medical Association Journal = Journal De L'Association Medicale Canadienne, 183(7), E391-402. https://doi.org/10.1503/cmaj.101860

Van Walraven, C., Dhalla, I. A., Bell, Chaim, Etchells, E., Stiell, I. G., Zarnke, K., . . . Forster, A. J. (2010). Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. CMAJ : Canadian Medical Association Journal = Journal De L'Association Medicale Canadienne, 182(6), 551–557. https://doi.org/10.1503/cmaj.091117

Van Walraven, C., Wong, J., & Forster, A. J. (2012). Derivation and validation of a diagnostic score based on case-mix groups to predict 30-day death or urgent readmission. Open Medicine, 6(3), e90-e100.

Yu, L., & Liu, H. (2003). Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. In (pp. 856–863).

Zhou, H., Della, P. R., Roberts, P., Goh, L., & Dhaliwal, S. S. (2016). Utility of models to predict 28-day or 30-day unplanned hospital readmissions: an updated systematic review. BMJ Open, 6(6), e011060. https://doi.org/10.1136/bmjopen-2016-011060